DOCUMENT RESUME

ED 317 562 TM 014 591

AUTHOR Seligman, Dee
TITLE A Look at Student Achievement from the School Dimension: Demythologizing Standardized Tests. Critical Issues in Student Achievement. Paper No. 3.
INSTITUTION Southwest Educational Development Lab., Austin, Tex.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 89
NOTE 29p.; For related documents, see TM 014 589-590.
AVAILABLE FROM Southwest Educational Development Laboratory, 211 E. 7th Street, Austin, TX 78701-3281. ($2.50 plus $1.50 shipping and handling).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *Educational Assessment; *Educational Improvement; Elementary Secondary Education; Sampling; *Standardized Tests; Test Bias; *Testing Problems; Test Interpretation; Test Use

ABSTRAC.

Issues emerging from a look at student achievement, which is defined in its usual school context as achievement on standardized tests, are addressed. A mythology about standardized testing has developed, in part because the metaphoric languages of medicine and business have been applied to education, defining it in terms of cure, efficiency, and productivity. Test bias, political implications, and contemporary learning theory can transform standardized test use into test abuse. Growing awareness of the dangers of exclusive dependence on standardized testing has resulted in many changes, including: (1) use of tests for screening and diagnosis; (2) innovative statewide assessment programs; (3) use of means of assessment that are authentic measures of what students need to gain; (4) use of large-scale tests in innovative ways; (5) goal planning as the emphasis for raising achievement; (6) new technology for testing; (7) increased awareness of fair testing practices; (8) development of the National Commission on Testing and Public Policy; (9) use of expectancy scores; (10) use of sophisticated sampling techniques; and (11) use of standardized tests as only one source of information. Standardized testing must come to be regarded, not as a cure for educational ills, but as a useful tool with much potential for misuse. (SLD)

CRITICAL ISSUES IN STUDENT ACHIEVEMENT

A Look at Student Achievement from the School Dimension: Demythologizing Standardized Tests

Paper Number 3
Spring 1989

Published by
SEDL
Southwest Educational Development Laboratory

Critical Issues Papers are published through Southwest Educational Development Laboratory's Programmatic Theme E - Facilitating Student Achievement in Reading, Writing, and Thinking, with Partnerships and Technology. SEDL, 211 E. 7th St., Austin, Texas 78701-3281, (512) 476-6861, SourceMail RD0260.

Sponsored by the
OERI
Office of Educational Research and Improvement

ACKNOWLEDGEMENT

My sincere appreciation to the following people for their review of this manuscript and helpful ideas in shaping it: Jo Ann Canales, Ida Jean Holman, Magdalena Rood, Kris Taylor, and David L. Williams, Jr. Their attentive and sensitive readings helped clarify and communicate the important points of this paper. The patience and talent of Penny Seago, Lonne Parent, and Pam Rowe in producing it were the other part of the team who made it possible.

Dee Seligman, Ph.D.
Training and Technical Assistance Associate
Resources for School Improvement
Southwest Educational Development Laboratory

A LOOK AT STUDENT ACHIEVEMENT FROM THE SCHOOL DIMENSION: DEMYTHOLOGIZING STANDARDIZED TESTS

by

Dee Seligman, Ph.D.

BACKGROUND

Southwest Educational Development Laboratory (SEDL) has been studying ways to facilitate student achievement in reading, writing, and thinking with partnerships and technology. In a series of critical issues papers SEDL has been looking at the nature of student achievement and the issues surrounding its facilitation in these content areas. SEDL staff has modeled an integrated framework which focuses on three dimensions: achievement from the perspective of the school, of the individual student, and the community. The first critical issues paper discussed the demographic, economic, and legislative background for educational reform. The second critical issues paper discussed the integrated model from which a broader definition of student achievement is drawn. It also looked at the impact on student achievement of the trend towards process-based instruction and of the trend towards considering the educational system in a socio-economic context.

This paper will address the many critical issues which emerge from a close look at student achievement defined in its usual school context as achievement on standardized tests. This paper addresses the metaphoric language of medicine and of business which informs our thought about education and through which a mythology about standardized testing has developed. By analysis of the operative metaphors, the paper looks at the underlying meaning assigned to standardized testing. The paper suggests that through such a process standardized testing can be demythologized and better understood. Seen from this new perspective standardized testing has limited, very useful functions which are potentially dangerous if we allow the metaphors to operate in a formulaic way. The paper provides a more informed understanding of the role for standardized testing through provision of a historical context, explanation of key terminology, consideration of three key issues of standardized tests, and, finally, by consideration of recent developments taking place in assessment.

THE METAPHOR OF MEDICINE

Once upon a time in the land of Wonderland, a prestigious national commission decided to change medical care in order to address the poor state of health care in its country:

> In response, a major hospital decided to institute performance measures of patient outcomes and to tie decisions on patient dismissal as well as doctors' salaries to those measures. The most widely used instrument for assessing health in Wonderland was a simple tool that produced a single score with proven reliability. That instrument, called a thermometer, had the added advantage of being easy to administer and record. . . .
>
> When the doctors discovered that their competence would be judged by how many of their patients had temperatures as measured by the thermometer as normal or below, some complained that it was not a comprehensive measure of health. Their complaints were dismissed as defensive and self-serving. The administrators, to insure that their efforts would not be subverted by recalcitrant doctors, then specified that subjective assessments of patient well-being would not be used in making decisions. Furthermore, any medicines or treatment tools not known to directly influence thermometer scores would no longer be purchased.
>
> After a year of operating under this new system, more patients were dismissed from the hospital with temperatures at or below normal. . . . Some years later, during the centennial Wonderland census, the census takers discovered that the population had declined dramatically and that mortality rates had increased. As people in Wonderland were wont to do, they shook their heads and sighed, "Curiouser and curiouser" (Suhor, 1985, p.635).

Such a tale is exquisitely appropriate to be told of the current situation in testing in American education. Our country frequently uses medical

analogies to describe the processes of education. Teachers are professionals who use scientifically developed "instruments" in order to "diagnose" their students' ills and "prescribe" appropriate "treatments." Educators speak of "at-risk" students, as if a fatal disease were imminent rather than a complex group of cognitive, social, and emotional factors that affect success in school or even completion of school. We attempt to find rational means by which to quantify what students have gained in school. Then the data are electronically manipulated, compared to where we want our students to be, and published on wall charts as a barometer of our national educational health.

This medical metaphor, which pervades the thinking about American education, might not be harmful except that it is taken dead seriously. The traditional concept of medicine as a process of curing disease, not of preventing it, with the help of science and technology might actually provide creative insights into the processes of education. For example, just as physicians look at antibiotics as the wonder drugs, educators search for cure-alls. Similarly just as physicians need to be watchful that technology not subsume the physician's mandated task of closely observing the patient, teachers need to use technology as a tool to assist them in observing and guiding student learning. Otherwise, educators can fall prey to the same dangerous trap that lurks in the sophisticated technology available to contemporary physicians:

> Technologies that improve accuracy, and centralized organizations that enhance efficiency and improve security, are essential factors in modern medicine. Yet accuracy, efficiency, and security are purchased at a high price when that price is impersonal medical care and undermining the physician's belief in his own medical powers. To be freer to develop his medical skills to their highest point, to increase what is despite these problems a positive balance of benefits over harms, today's physician must rebel. He can use his strongest weapon—a refusal to accept bondage to any one technique, no matter how useful it may be in a particular instance. He must regard them all with detachment as mere tools, to be chosen as necessary for a particular task. He must accept the patient as a human

> being and regain and reassert high faith in his own medical judgement. (Reiser as quoted in Haney and Madaus, 1989, p.687)

The medical metaphor for education, thus, has its usefulness, but it is inherently dangerous if reduced to a simple clichéd formula. Medicine is both an art and a science. Education can mimic the science of medicine but, in doing so, must recognize there are areas in which the science has inherent restrictions. A standardized test can only give limited amounts of information like a thermometer and cannot be used rigidly for all decisions or even for all measurements.

We now recognize that medicine should be preventative as well as restorative, although the processes of proactive medicine and proactive education are extremely complex. What works for one patient/student will not work for another. We need a much more complex view of the processes inherent in medicine if we are to apply them analogously to education. This analysis of the pervasive medical metaphor reaches to the heart of the body of demands for educational reform and effective teaching. Until American educators have a clear sense of what Wiggins (1989) calls our "national intellectual fitness," they will neither be able to develop appropriate mechanisms for measuring that fitness nor effective strategies for coaching towards it. Assessment, whether in the form of standardized testing, performance assessment, or teacher-designed classroom tests, is not a dessert added on at the end of the instructional meal. Rather it is the activating force of the whole educational process. Wiggins speaks of "authentic tests," those tests that are true tests of ability because their criteria for success are both known and valued. Such tests contrast with traditional standardized tests which are single tests documenting an estimated performance on a carefully narrowed range of skills in a particular slice of time under a given set of circumstances (Wiggins, 1989). He suggests that:

> The redesign of testing is thus linked to the restructuring of schools. The restructuring must be built around intellectual standards, however, not just around issues involving governance, as has too often been the case so far. Authentic restructuring depends on continually asking a series of ques-

tions: What new methods, materials, and schedules are required to test and teach habits of mind? What structures, incentives, and policies will insure that a school's standards will be known, reflected in teaching and test design, coherent schoolwide, and high enough but still reachable by most students? Who will monitor for teacher's failure to comply? And what response to such failure is appropriate? How schools frame diploma requirements, how the schedule supports a school's aims, how job descriptions are written, how hiring is carried out, how syllabi and exams are designed, how the grading system reinforces standards, and how teachers police themselves are all inseparable from the reform of assessment. . . . Only such a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness. As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching our intellectual standards, student performance, teaching, and our thinking and discussion about assessment will remain flaccid and uninspired (p.712).

. . . Only such a humane and intellectually valid approach to evaluation can help us insure progress toward national intellectual fitness. As long as we hold simplistic monitoring tests to be adequate models of and incentives for reaching our intellectual standards, student performance, teaching, and our thinking and discussion about assessment will remain flaccid and uninspired (Wiggins, 1989).

HISTORY OF STANDARDIZED TESTING AND THE BUSINESS METAPHOR

The history of standardized testing reveals much about the business mythology surrounding the educational process. During World War I, the U.S. Army was the first to use a test of intelligence to gain a range of information quickly and efficiently. While the administration could hardly be called standardized, the test allowed decisions to be made about the recruits so that they could be placed in appropriate duties. One consequence of testing at that time was the permanent labeling of recruits according to tested mental abilities.

During the years 1911-1916 as America became industrialized, there was a "school efficiency" movement afoot not unlike that of contemporary times. Wiggins (1989) discusses the analogies drawn between Frederick Taylor's management principles, used to improve factory production, and the principles used to improve schools. Higher standards were demanded of the schools, a demand which was usually translated into increasing workloads for teachers and students. As the school population grew with the influx of immigrants, then current views that intelligence was related to social class and heredity, came into play. Society

was stratified economically and socially. The tests were used at least in part as sorting mechanisms. As one superintendent put it: "The results of a few well-planned tests would carry more weight with the businessman and parents than all the psychology in the world" (Callahan, 1962 as quoted by Wiggins, 1989).

After World War II the prevailing metaphor for education continued to be a business model, still essentially a factory model. The efficiency and productivity of American business became the standard of reference and is, astonishingly enough, still the operative one. The National Defense Education Act of 1958 required testing in order to establish this bottom line, which was associated with a Defense Department mentality of objective-based management. (Madaus, 1989). Other federal legislation such as the Elementary and Secondary Education Act of 1965, the Rehabilitation Act of 1973, and P.L. 94-142 (the Education for All Handicapped Children Act) also required testing (Madaus, 1989). Distrust of educators and the perception that schools were not doing a good job greased the rails for the engines of the American educational testing program. By laying down clearly standardized tracks, the educational community persuaded the American public that it could see where the schools were going and how fast they were getting there. There has been little movement away from this business model even in the 1980's. We still talk about "choice" in the public schools as if students could actually shop to choose their courses at different schools in the way that many haunt department stores for bargains.

In the 1970's the use of standardized testing became part of the national focus on literacy, in part initiated by Admiral Hyman Rickover's disgust with Navy recruits' inability to read or write. What is today often described as "accountability" needs at the state and national levels really began with a military sense of vulnerability if our troops couldn't understand their training manuals. Although the American military's attention to illiteracy has been a strong force in the emphasis on testing, there are some differences between its concern during World War I and in the 1970's. During World War I the military used intelligence testing as a way of sorting those whose intelligence, it was thought at that time, was not changeable. In the 1970's the emphasis had changed to a focus on imple-

menting increased literacy with the assumption that aptitude could change.

Salganik (1985) also delineates other causes for the popularity of large-scale standardized testing in schools. "Output controls," mechanisms for measuring clearly defined goals, became highly desirable in the late 1970's after the failure of attempts to make schools accountable through "process controls," that is, breaking down the educational process into a rational series of steps and objectives. Instead the public, many of whom felt politically disenfranchised, turned to the harder data of measurable goals since they distrusted educators. Test results seemed more objective and reliable (Salganik, 1985). Furthermore, standardized testing offered a convincing political tool for administrators and legislators that "scientifically" sound, technical protocols were being followed, not decisions based on personal, unreliable judgment. Indeed, as Salganik pointed out four years ago, the public has become increasingly comfortable with the idea of standardized testing as a normal occurrence. Policy questions about testing have been diverted into technical problems, such as what is the cutoff score for a minimum competency test, rather than whether that minimum competency test provides adequate assurance that intellectual health is being developed and sustained.

TERMINOLOGY/ CONCEPT OF STANDARDIZED TESTING

Standardized testing includes both norm-referenced and criterion-referenced testing, either of which may have norm-referenced interpretation of test scores.[1] The most basic distinction in testing is between standardized and non-standardized tests. A standardized test consists of standardized questions, administered in standardized conditions, and scored through uniform procedures; such a test is constructed by experts in the subject matter. The term standardized doesn't necessarily indicate that the test measures what should be taught, although such tests do typically measure the common essence of the "national curriculum."

There are basically two types of standardized tests: norm-referenced or criterion-referenced. Norm-referenced tests are used to rate the students' performance along an established continuum of objectives, while criterion-referenced tests are those which establish mastery or non-mastery of students on a selection of objectives. If the tests are norm-referenced, they refer to the average score of a clearly

specified norm group, whose scores were gathered at a particular point in time. The norm is the average for that particular group at that particulart point in time, but more or less than 50% of current students, such as current fifth graders, could be above the point which defined the 50th percentile for a particular norm group for a particular year. Norm-referenced standardized tests add meaning to their scoring by comparing it to the scores of students in the reference (norm) group.

Recently the Cannell report (1987) charged that since all state scores at the elementary level in all fifty states were above the national norm on the six major nationally normed tests, the information was not valid. Such a report misunderstands the nature of norming, since if students in 1989 are compared to a 1984 norm group it is entirely possible for more than 50 percent of them to be above the median (50th percentile). The other cause of such an effect is the damaging influence of teaching too closely to the test, i.e., restricting curriculum to test objectives or specific material covered by the test. A third possible cause, certainly more suspect, is that certain groups, such as learning disabled or Limited English Proficiency students, were intentionally excluded from the norm group. However, if the norm group was not very current, a fairly common occurrence, it would be entirely possible for most students to score above the 50th percentile.

The other concept that is important to an understanding of standardized testing is types of validity. One can only talk about the validity of the inference that is made from the score, not the validity of the test itself. One of the most fundamental forms of validity is content validity, that is the degree to which the sample of items on the test represent some defined domain of content. Thus, one can only talk about the validity of the test itself. Since educators cannot teach to the total domain, they must sample from it and make inferences about the total domain. If there is too much teaching to the test instead of to the total domain, the inference to the domain is no longer valid. Thus, it is possible to increase test scores by tying curriculum to the "national curriculum" of the major achievement tests, but not really increase the actual standards of learning. Providing evidence for accountability based on test scores is considered inappropriate by many measurement experts because it is very difficult to establish causal relationships between

the test data and who or what is responsible for that data. The input variables are extremely complex and can not always be directly attributed to teachers or their instruction.

Content validity does not refer to whether the test content has actually been covered in curricular materials; that type of validity is called instructional/curricular validity. Content validity merely refers to the degree to which the test samples a domain, whether or not students had ever had instruction in a specific subject. For example, if test items adequately sampled the defined domain of Spanish, such a test would have content validity, whether or not students had actually had instruction in Spanish.

ISSUES ASSOCIATED WITH TESTING

In recent years complex issues related to test bias have emerged. Many groups, like the Center for Women Policy Studies, believe that pivotal tests like the Scholastic Aptitude Test, the most widely used college entrance exam, are gender-biased, in favor of boys ("16 Percent Biased," 1989). This Center reports that boys do better on mathematics questions by 44 points and on verbal questions by 14 points in part because the questions are biased towards subjects which interest boys, such as computation of a basketball team's win/loss record, or language indicative of boys' interests, such as the analogy "mercenary is to soldier as hack is to writer." Other researchers suggest that the cause of bias in standardized tests relates to language skills, which may be weak among bilingual, English as a Second Language, or minority students since the tests are written in stylized language rather than in common vocabulary. In addition, some words may have specialized meaning for particular groups, such as "environment" meaning "home" or "people" for Black students and meaning "air," "clean" or "earth" for Whites. A third explanation for bias relates to different cultural experiences. For example, the WISC-R IQ test asks the appropriate remedy for a cut finger. Although the correct answer for scoring purposes is to put a Bandaid on it, inner-city children usually respond "cry, bleed, or suck on it," (Medina and Neil, 1988).

In response to such assertions of gender, language, and culture bias, the Educational Testing Service, which develops and monitors the widely-used Scholastic

Aptitude Test (SAT), argues that this exam is not biased and actually predicts first-year college grades better for women than it does for men and better for Black and Hispanic students than high school grade-point averages. ("SAT Predicts," 1989).

The issue of bias is a very real one for high-stakes testing, that is, tests whose results are used for promotion, placement, and other pivotal decisions in an individual's school or teaching career. The response of testing organizations such as the Educational Testing Service (ETS) might best be suggested by one of its Distinguished Research Scientists, William Angoff. He explains that a difference in average scores between, for example, White and Black students, in and of itself does not constitute evidence of bias in the test. Most testing experts believe that the disparity reflects the "result of long-standing bias in our society that has deprived Blacks of the level of quality in education typically enjoyed by mainstream Whites" (Angoff, 1987, p. 2). He explains that one way of looking at equity, the absence of bias, is that what is equitable, fair, and unbiased in some situations is inequitable, unfair, and biased in other situations. He also urges consideration of the difference between fairness, based on the use of a biased test, with bias, the simple outcome of an analysis.

Angoff explains that individual items on tests are analyzed through a specific process. Two groups, such as one of White students and one of Blacks, are matched in terms of their ability on some measure such as a test score, usually the test itself that is under consideration. Although psychometricians acknowledge that it is philosophically troubling to compare groups according to the very test which is being analyzed, they argue that there is no better criterion available for matching. They do so by matching the two groups, looking at the way in which the results of each group plot out on a graph, and then look for individual test items whose answers fall at an unusually large distance from the central tendency of this plot. Thus, the items which fall into a plotted pattern become the "criterion" against which individual items are identified as "biased" because they diverge markedly from that criterion.

Angoff (1987) demurs that finding such differences still does not indicate bias since one wonders whether the matched groups are truly representative of their respective populations. He notes other reasons for statistical differences, such as items based on regional experiences (e.g. "name three parks in Manhattan") or items which focus on known disparities in performance (such as the known disproportionate difficulty of Black middle-school students with percentages).

The basic dilemma here resides in the question, wha. is the locus of test bics, the items of the test or the basic inequality in our society? And do we have the wit to tell the difference? (Angoff, 1987)

Thus, he argues the three basic issues concerning bias are: (1) distinguishing conceptually between bias resulting from unequal opportunity in society, and bias resulting from the structure and content of particular test items; (2) understanding that bias in tests exists only in the context of the purpose and use of the test and in the type of inferences drawn from the test scores; (3) finding good analytic procedures that are philosophically sound for detecting bias. For these reasons, he concludes that disparities in performance between two groups of individuals, even when matched on a suitable criterion are called "differential item functioning"or "differential item difficulty," n bias (Angoff, 1987). He acknowledges that the relationship between aptitude and achievement is difficult for many people and is embedded in controversy about the nature of modifiability of intelligence. Many people confuse the construct of aptitude with the tests used to measure them. He believes that we should distinguish between the instruments of testing and the constructs of aptitude and achievement, since the constructs still show validity.

Willie (1985) pushes the bias argument even farther by suggesting that as a free and pluralistic society we need to depend much less on standardized tests. He assumes that ethnic groups, based on their backgrounds, do have cultural diversity and that we should cultivate this diversi instead of insisting on uniform measurement scales. He believes that applying a common criterion to everyone in the sphere of education guarantees injustice because of these cultural differences. Depending too heavily on standardized tests negates the great benefit to higher education of finding students whose experiences and abilities will enrich the learning environment and whose personal characteristics will later contribute to society. Clearly, questions of bias are a potent danger when the use of standardized testing moves to high-stakes decisions.

A second area of concern about standardized testing is the extent to which testing has become a political football. Psychometricians argue that the question of who should be designing the objectives of education, State Departments of Education and their academic counterparts in colleges of education or the individual teachers, is not a measurement issue, but a philosophy of education question. In fact, psychometricians do engage in policy issues, evidenced by a panel of experts' recommendation in a study commissioned for the Texas Education Agency (TEA). They argued against the idea of student achievement scores being used for career ladder decisions for teachers. They advised the TEA that the "problems associated with using student achievement data to make decisions about teacher career ladders outweigh the benefits," especially the difficulty of making valid inferences about teacher effectiveness at the classroom level (TEA, 1988).

There is extensive use of standardized test scores to hold states and school districts accountable for the quality of education. The U.S. Department of Education has published a wall chart for the last six years documenting test results of the SAT and ACT, along with graduation rates, teacher salaries, and education spending ("Cavazos Hopes," 1989). However, the inferences drawn based on these scores are by no means clear. Steelman and Powell (1985) suggest that the state-by-state variations need to be corrected for the percentage of students taking the exam, and the distribution of test takers by sex, race, and socioeconomic status. The corrected state scores rank the states very differently. Furthermore, Steelman and Powell suggest that some of the causal explanations offered for higher scores, such as an increase in academic course work, are not borne out by the research. Research shows that other causes that are often overlooked, such as the amount of money spent on students in different states, seem more directly related to these corrected SAT scores. They conclude that such test scores can be manipulated in many different ways that have very little to do with the quality of education but have a lot to do with our dependency on SAT scores.

One cannot separate the test from its use. The issue is that the kind of criterion used, the nature of the assessment itself, defines the nature of what is taught, how it is taught, and how it is learned. While many deride teaching to the test, that fact

is that the tests are meant to drive the curriculum to clarified instructional objectives, demand consistency and quality in teaching instruction, and define for the public the minimum and maximum competencies for which a given state is responsible. Madaus (1985) speaks of standardized tests as the "linchpins of policy." This debate is really one of educational goals and policies, but the tests themselves predefine the policies. Thus, Wiggins (1989a) suggests that we do need to teach to the test but to "standard-setting tests" that accurately reflect those habits and skills of reading, writing, questioning, speaking, and listening that we want our students to learn. Instead, we should teach to what he calls "authentic tests."

The common thread that links these human perspectives on the meaning of test scores is the use of these scores as administrative mechanisms by which to implement one or another policy. In each case, testing as an administrative device has become the linchpin of policy (Madaus, 1985).

The extent to which school districts encourage teachers to teach to the test was discussed at length at a conference on assessment held in Summer 1988 by the Colorado Department of Education and the Education Commission of the States (Pipho, 1988). There was a lack of consensus on what constitutes appropriate or inappropriate activity vis-a-vis testing. Curricular alignment between a school district's objectives and state-mandated objectives, even to the level of correlations between textbook units, supplementary materials, and the state's test data, seemed appropriate to some districts. They believe that realigning the curriculum is the preferable approach since districts with poor test performance can only do one of two things: provide remedial help for students after the fact, or realign the curriculum.

In fact, increasingly states are using test scores on achievement and competency tests as one of the pieces of data used for accountability in making policy decisions which may include rewards or sanctions to the school districts. Forty five states do collect data on achievement test scores and twenty-five states do have policy links to accountability data (*Creating Responsible Systems*, 1988). For example, in Arkansas legislation requires schools to meet certain accreditation standards, which include adequate performance on the Minimum Proficiency Test (MPT) or face the threat of consolidation. In South Carolina data on student achievement on norm-referenced tests of basic skills and State-developed tests of mastery are used for decisions on withholding of state funds (in extreme cases) or

monetary rewards for greater-than-expected gains in student achievement (*Creating Responsible Systems*, 1988). South Carolina's governor and key lawmakers are currently proposing that schools with high marks on the state's standardized test, as compared with other schools whose students have a comparable socioeconomic profile, would be released from compliance with such state regulations as class scheduling, class structure, and staffing (Flax, 1989). The Texas legislature is currently considering a bill proposed by the Governor to reward schools financially for scholastic gains, one criterion of which is student achievement. In New Jersey the state can fire district administrators and disband school boards if the state deems the district "educationally bankrupt," a decision based on 51 indicators, one of which is student performance on State tests (*Creating Responsible Systems*, 1988).

Those states that have minimum competency testing are in fact using these tests to drive instruction. As Popham (1985) explains in reviewing the results of three states and one city with carefully crafted minimum competency programs, all four had one feature in common: measurement was used as a catalyst to improve instruction and clarify instructional targets. Isolation of the skills for which students will be held responsible and communication of expectations to teachers and administrators were significant components of such testing programs. While one can recognize the urgency of the states to find successful means to help students learn, there is considerable discussion about whether minimum competency testing is the most effective vehicle for doing it.

The new literacy of thoughtfulness calls for a quite different technology of teaching and testing...it is about the making of meaning, not just the receving of it. Thoughtfulness is a constructive, not a passive, undertaking (Brown, 1989).

A third area of issue concerns whether the standardized tests adequately encapsulate what is known about learning, an area where many changes of understanding have occurred in the last decade. Educators no longer assume students are passive "blank-slates" into which knowledge is poured and then demonstrated upon teacher demand. We look at learning as more of a constructive process, and knowledge as something that is socially and personally derived through social interaction, implementation and testing of ideas, and personal discovery rather than as formulated, hierarchical truths that can be simplified into accessible language and taught. Most of the time we sift through the various types of knowl-

edge that we already have and try to apply or evaluate with it rather than simply derive it (Brown, 1989). Thus, assessment must reflect the active nature of learning and not simply short-term memory over inert knowledge. Since the demands on adults as learners change almost daily, educators know that learning how to learn is a significant part of what is now called literacy. These new conceptions of learning are implicit in the programs called for by the Coalition of Essential Schools, by the extensive work in critical thinking, by the whole language movement, and by the new focus on collaborative learning for staff development.

The issues surrounding the nature of learning become significant in the context of standardized testing because most tests, as they are currently designed, can only assess very specific objectives built around discrete skills. The tests assume a linear scale on which students develop from simpler to more complex learning that can be symbolized by a single score (Medina and Neill, 1988). In fact, we know that children often do learn the use of pronouns before they learn an extensive noun vocabulary, and some concepts of physics and mathematics are understood long before a child has the vocabulary with which to express the ideas. We also know that a unitary score does not reflect the complexity of human learning or intelligence, which is multi-dimensional and may be highly developed in one area but not in others. Gardner (1985) hypothesizes seven different intelligences which are distinct and independent.

The idea that assessment should reflect current understanding of learning is exemplified in the area of reading. We now understand reading as a process in which topic familiarity is very important to reading comprehension. We know that good readers know how to use inferences to extract information and to interpret it. We know that reading is a process in which skilled readers use metacognitive strategies to get meaning from the reading and to monitor their own misunderstandings or lack of understandings. We also now understand that reading is highly influenced by positive attitudes, fluency with written material based on wide literacy experiences, and the ability to restructure information based on integration with the individual reader's prior experience (Valencia, Pearson, Peters,

Wixson, 1989). Yet standardized tests, for the most part, are used to assess a wide variety of skills and subskills in short passages taken out of context. Little attention has been focused on the metacognitive strategies used by readers, on the reader's literacy experiences at school or at home, or on their familiarity with the topics (Valencia, Pearson, Peters, and Wixson, 1989; *Assessing Reading in Illinois*, 1989). Until quite recently no statewide standardized tests used the current research on reading and learning to assess reading. Michigan and Illinois are pioneer states in creating reading tests that reflect the current research.

NEW DEVELOPMENTS IN TESTING

The three areas of concern raised about standardized testing—bias, political implications, and contemporary learning theory—all can transform testing use into abuse, depending on how the data collected from standardized tests are disseminated and for what purposes. Because of a growing awareness of the dangers for the educational process of exclusive dependence on standardized testing, many changes are occurring, including:

•*Use of tests for low-stakes testing*, e.g., valid developmental screening to identify high-risk children and help teachers diagnose educational needs.

•*Innovative state-wide assessment programs* such as the teacher-developed California Assessment Program's Writing Assessment which measures, through matrix sampling, eight types of writing at grades six, eight, and twelve (grade three in 1990-91). Students exhibit their performance by writing in one of eight different forms, such as autobiographical incident, report of information, problem solution, evaluation, eyewitness memoir, etc. Matrix sampling means that the scores cannot be computed for individual pupils, since every student's test is in fact a randomly assigned subpart of a larger test. Since a broad content domain is being tested, it is less possible to teach to the test through instructing in a narrow range of skills. The California Assessment Program is unique because it includes a wide variety of writing types and encourages teaching a variety of literary genres. Higher level thinking is implicit in the types of prompts, which use not only curriculum-based information, but experience-based ideas and literature-based memories. These tests encourage writing across the curriculum and are

achievement tests, not minimal competency tests (*California Writing Assessment Handbook, Grade Eight*, 1986).

•*Use of performance assessment, productions, exhibitions of mastery, portfolios, teacher inventories, and other means* of measuring student growth and learning that are authentic representations of what we want students to gain from their educational experiences (Wiggins, 1989; Archbald and Newmann, 1988). Some states, such as Connecticut, are moving to statewide student assessment based on performance. They are developing a Common Core of Learning for educated high school graduates that, in 1990-91, will focus on math and science. Their assessment will use portfolios, simulations, exhibitions, and projects to look at how students integrate their knowledge, skills, and attitudes in active and collaborative learning situations. The schools affiliated with the Coalition of Essential Schools also are fostering performance assessments (Shepard, 1989; Wiggins, 1989a; Wiggins, 1989b).

•*Use of large-scale tests in innovative ways.* For example, for the last six years the Ohio Board of Regents has authorized an English writing sample test for high school juniors that is holistically scored both by high school composition and college English freshman writing teachers (Pine, 1985). The collaborative scoring of the test, with comments from both groups of teachers, gives students feedback on their writing, encouraging their demand for more writing instruction. Equally interesting, the use of this test and its scoring procedures have had a dramatic effect on staff development, since the high school teachers have a clearer understanding of what is expected by their counterparts in higher education. It has had a significant effect on freshman writing at Youngstown State University, as well as on the teachers' collaborative work as writers with their students.[2]

Ohio also has an Early Math Placement Test, which is administered to high school juniors to predict their placement in math courses at their intended Ohio college or university in an intended field of study. This information is provided to the students and to their high schools to encourage students to continue taking senior-level math courses. The data are neither published nor are they actually used for

placement. Results include the development of a numerically-based problem solving course to bridge the movement from numbers to symbols for math-weak seniors. Ohio now makes the same course available for seventh and eighth graders. This inexpensive test has had a significant effect on raising the level of math placement at Ohio State University.[3]

•*Goal planning* as the emphasis for raising student achievement. School districts such as Alamo Heights Independent School District in San Antonio, Texas have set goals for higher test scores on standardized tests. Their emphasis is on specific steps for reaching those goals, such as reviewing the whole curriculum regularly, adding reading courses in seventh and eighth grades, providing a junior high writing laboratory, teaching vocabulary through the twelfth grade, refining graduation requirements, and increasing the number of Advanced Placement courses (Pine, 1985).

•*New technology breakthroughs in testing*, particularly adaptive computer tests that lead students from one level of difficulty to another, teaching and testing at the same time (Brown, 1989). The Educational Testing System (ETS) also is exploring ways to use artificial intelligence technology to score free-response test questions (i.e., open-ended questions that resemble real-life tasks). At this time, ETS researchers are looking at ways to grade free-response questions on the Advanced Placement examination in computer science. Their work suggests that computers can, under specific limitations, grade constructed responses, especially the more well-formed ones, as well as human experts grade them ("Researchers Believe Artificial Intelligence," 1988).

In addition, ETS is developing computer-based testing systems that will measure and track a student's growing knowledge and skill in a subject ("Computer Simulations," 1988). Called Mastery Assessment Systems, the approach involves computer simulations that engage students in using higher-order thinking skills such as planning, analysis, troubleshooting, and hypothesizing. For example, mastery assessment in science might depict a variety of realistic electrical components on the computer screen, and ask the student to select and assemble the components

into a working circuit. Thus, the process is very different from a multiple-choice test that asks, "Which of the following is a complete circuit?" In the typical multiple-choice test the sequences of test items bear no particular relationship to each other, and the facts that are measured are arbitrary and not necessarily related.

In the Mastery Assessment process the computer's assessment includes rules for interpreting the student's progress through various steps in order to better interpret what is going on in the student's mind. Although teachers have been doing this for a long time, the challenge now is to give the computer rules for interpreting this evidence. The student's score is not a number, but a two-dimensional mastery map conveying the boundary between what the student does and does not know. Associated databases for further instruction and tutorials will be linked in to the assessment process. In this way, the boundary between instruction and assessment will be blurred in future on-line testing ("Computer Simulations," 1988). As a result, Madaus (1989) asserts, use of technology will lead to a decrease in multiple-choice tests and an increase in tests that allow students to produce answers.

•*Increased testing company awareness of what constitutes fair testing practices.* The Code of Fair Testing Practices in Education, a cooperative effort of the major testing companies and measurement-oriented professional organizations, explains the roles of test developers and test users, the rights of test users, the potential misuse of valid tests, the necessity of eliminating all bias whenever possible, and the availability of information about the scoring of the test and control of that scoring (*Code of Fair Testing Practices in Education*, 1988).

•*Development of the National Commission on Testing and Public Policy*, which began at the University of California at Berkeley, to look at the social, economic, and political issues surrounding testing. Its members include educational leaders, business leaders, political leaders, civil rights leaders, academicians, and the former President of the College Board. The commission will make recommendations to policy makers about the use of standardized tests in the educational process and will set the context for future testing policies.[4]

•*Use of expectancy scores*, which are a comparison between the score on a standardized test and some other measure such as an aptitude test. Mehrens (1989) suggests that by giving two tests, an aptitude and an achievement test, published by the same company in the same year, using the same norms, one can derive a much fairer measure of what can be expected from a student. Giving such a configuration twice during the elementary years would, according to advocates, yield much more useful information with which to evaluate educational programs. Accountability is less of an issue with use of an expectancy score.

•*Use of more sophisticated sampling techniques*, such as the matrix sampling used in the California Assessment Program and the National Assessment of Educational Progress. Such a sampling requires that there be a new test each time, which has a large item bank from which different objectives are sampled each year. Although this method yields no information about individual students, it avoids teaching directly to the test and provides school- and district-level information.

•*Use of standardized testing as one source of information* among many other sources, not as the sole measure through which students, teachers, and administrators are held accountable by the public or the state.

CONCLUSION

Many changes are occurring within this field of testing, but the existence of standardized tests and some form of dependence on them is likely to continue in the foreseeable future. It is clear that standardized testing offers certain benefits. Through testing educators can strengthen and define their instructional objectives. The tests can provide feedback to teachers, administrators, and the public about what is being taught. Some tests can provide diagnostic information useful on an individual basis. If data are used correctly, administrators have helpful information to make decisions in such areas as developmental screening and placement in special classes. Standardized testing can encourage better and more varied teaching, can assist school districts in long-range goal planning, and can drive instruction to cover the minimum competencies.

There are, however, many liabilities in dependence on standardized testing, some of which the public and even some educators are unaware. There is some learning that can't be measured by standardized tests, but that should not negate the value of these objectives. It is increasingly easy for teachers to teach to the tests, and there is enormous pressure at every level, from local school district through the federal level, to do so in order to improve test scores.

There are many ways to manipulate the data derived from standardized testing, including changing the norm by the inclusion or exclusion of specific populations, such as a school with a high drop-out rate having high scores since these students usually score poorly on tests. Standardized testing only provides a small sample of what actually is going on in the classroom or in the learning environment. Indeed fewer than 20 percent of more than 350 policymakers and practitioners in four states interviewed between 1985 and 1987 by the Center for Policy Research in Education (CPRE) felt that test scores were the most important source of information about a district or school. Half the superintendents and principals interviewed did not even mention State test scores as among the major sources of information they used to determine how well their districts or schools were doing (Pine, 1988).

The public misunderstands the significance of test scores. With state-by-state and district-by-district comparisons published in the media, the public believes they have a scientific hold on the quantity of learning taking place in American schools. There is a wide-spread but mistaken belief that high scores are synonymous with good education.

A further liability is that the curriculum can be decided by what is included or excluded from major standardized tests, which are developed by private industry without governmental restrictions, rather than by professional teachers and administrators.

The issues are both complex and many, but they may be summarized in three areas:

- bias and the concomitant implications for high stakes decisions for minority groups;
- political uses of testing by legislators and administrators;
- incongruence of current standardized tests with contemporary learning theories.

Understanding the history, the primary concepts behind standardized testing, the strengths and weaknesses, and the issues should provide a clearer context for taking this curious tool out of Wonderland. The current use of standardized testing is part of a larger mythology in American education. In fact we know that speaking of achievement as if it could be adequately measured solely by current standardized tests is a very narrow and inaccurate way of defining achievement. Either we can be deceived into thinking that students are learning more than they actually are, i.e., be overconfident, or we can be disillusioned that they are learning very little, i.e., be underconfident, when we are not measuring their real learning and thinking. We need to look at standardized testing not as a magical cure for American education, but as a potentially dangerous, sometimes useful tool, which, if properly controlled, could yield helpful information that can corroborate other sources of information about student achievement.

ENDNOTES

[1]This information summarizes material presented by Dr. William Mehrens, Distinguished Professor from Michigan State University, and Dr. Glynn Ligon, Executive Director, Department of Management Information, Austin Independent School District, at meetings for SEDL's Theme E partners. Their presentations, dealing with test construction, validation, and application, are explained in greater detail in two papers which are available upon request from SEDL.

[2]Based on a conversation with Dr. Elaine Hairston, Vice-Chancellor for Academic and Special Programs, Ohio Board of Regents, May 1989. Further information on impact of this program is available from Dr. Grace Murphy, Youngstown State University, Department of English.

[3]Elaine Hairston, Ohio Board of Regents, conversation in May 1989. For further information on secondary curriculum developed as a result of the Early Math Placement Test, contact Professor Burt Waits, Ohio State University, Department of Mathematics.

[4]The National Commission on Testing and Public Policy, University of California, Office of the Dean, School of Education, Tolman Hall, Berkeley, CA 94720 or Dr. George Madaus, Executive Director, National Commission on Testing and Public Policy, Boston College, McGuinn Hall 531, Chestnut Hill, MA 02167 can provide further information on the Commission and its publications.

REFERENCES

Angoff, W. (1987). *Philosophical issues of current interest to measurement theorists.* Princeton, NJ: Educational Testing Service.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* Reston, Virginia: National Association of Secondary School Principals.

Brown, R. (1989). Testing and thoughtfulness. *Educational Leadership, 46(7),* 31-35.

California State Department of Education. (1986). *Writing assessment handbook - Grade Eight.* Sacremento, CA: CSDE.

Callahan, R. (1962). *Education and the cult of efficiency.* Chicago, IL: University of Chicago Press.

Cannell, J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average.* Daniels, WV: Friends for Education, Inc.

Cavazos hopes to stir up stagnant educational waters. (1989, May 4). *Education Daily,* p. 3.

Code of fair testing practices in education. (1988) Washington, DC: Joint Committee on Testing Practices.

Computer simulations offer realistic assessment of higher order thinking skills. (1988, Fall). *ETS Developments,* pp. 8-9.

Flax, E. (1989, February 25). South Carolina considering "flexibility" for high-scoring high schools. *Education Week*, p. 1.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York, NY: Basic Books.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan, 70*(9), 683-687.

Illinois State Board of Education. (1988). *Assessing reading in Illinois.*

Ligon, G. (1989). *Test construction, validation, and interpretation.* Unpublished paper. Austin, TX: Southwest Educational Development Laboratory.

Madaus, G. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66(9),* 611-617.

Madaus, G. (1989). New ways of thinking about testing. *Phi Delta Kappan, 70(9),* 642-645.

Medina, N., & Neill, D. M. (1988). *Fallout from the testing explosion: How 100 million standardized exams undermine equity and excellence in America's public schools.* Cambridge, MA: FairTest.

Medina, N., & Neill, D. M. (1989). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 70(9),* 688-696.

Mehrens, W. (1989). *Achievement test construction, validation and application: The good, the bad and the ugly.* Unpublished paper. Austin, TX: Southwest Educational Development Laboratory.

Office of Educational Research and Improvement. (1988). *Creating responsible and responsive accountability systems.* Office of Education Publicaton No. PIP 88-808. Washington, DC: U.S. Department of Education.

Pine, P. (1985). *Raising standards in schools: Problems and solutions.* Arlington, VA: American Association of School Administrators.

Popham, J., Cruse, K., Rankin, S., Sandifer, P., & Williams, P. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66(9),* 628-634.

Reiser, S. (1979). *Medicine and the reign of technology.* New York, NY: Cambridge University Press.

Researchers believe artificial intelligence technology may be employed to score free-response test questions. (1988, Winter). *ETS Developments,* pp. 4-5.

Salganik, L. (1985). Why testing reforms are so popular and how they are changing education. *Phi Delta Kappan, 66(9),* 607-610.

SAT predicts college grades well, ETS study, says, but claims disputed. (1989, April 28). *Education Daily*, p. 8.

16 percent of SAT questions biased, researcher says. (1989, April 13). *Education Daily*, p. 3.

Steelman, L., & Powell, B. (1985). Appraising the implications of the SAT for educational policy. *Phi Delta Kappan, 66(9)*, 606-606.

Suhor, C. (1985). Objective tests and writing samples: How do they affect instruction in composition? *Phi Delta Kappan, 66(9)*, 635-639.

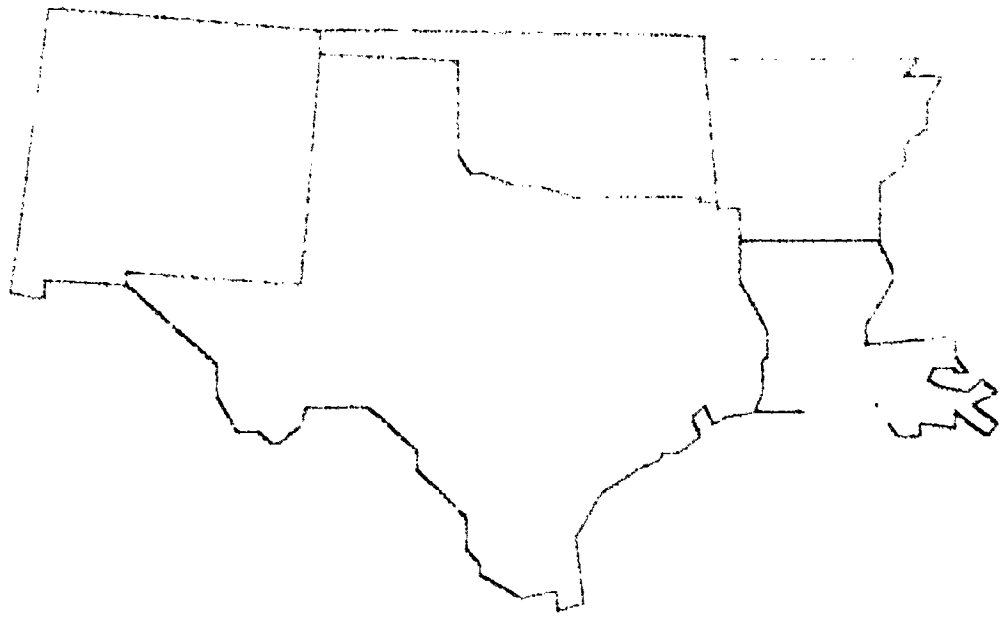
Texas Education Agency. (1988). A study to determine the most effective means of implementing career ladder level assignments that are made on the basis of student achievement in addition to other bases required by law. Austin, TX: TEA.

Valencia, S., Pearson, D., Peters, C., & Wixson, K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership, 46(7)*, 57-64.

Willie, C. (1985). The problem of standardized testing in a free and pluralistic society. *Phi Delta Kappan, 66(9)*, 626-628.

Wiggins, G. (1989a). Teaching to the (authentic) test. *Educational Leadership, 46(7)*, 41-50.

Wiggins, G. (1989b). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70(9)*, 701-713.



Southwest Educational Development Laboratory
211 E. 7th St.
Austin Texas 78701-3281
(512) 476-6861
SourceMail RD0260

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under Contract Number 400-86-0008. The opinions expressed in this publication do not necessarily reflect the positions or policies of the U.S. Department of Education or any other agency of the U.S. Government.